EXAMINER'S SEARCH NOTES CDT 5/24/05

US006785671B1

(12) **United States Patent**
Bailey et al.

(10) **Patent No.: US 6,785,671 B1**
(45) **Date of Patent: Aug. 31, 2004**

(54) **SYSTEM AND METHOD FOR LOCATING WEB-BASED PRODUCT OFFERINGS**

(75) Inventors: **David R. Bailey**, Lake Forest Park, WA (US); **Todd J. Feldman**, Seattle, WA (US); **Anand Rajaraman**, Seattle, WA (US)

(73) Assignee: **Amazon.com, Inc.**, Seattle, WA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/528,138**

(22) Filed: **Mar. 17, 2000**

Related U.S. Application Data

(60) Provisional application No. 60/169,570, filed on Dec. 8, 1999.

(51) **Int. Cl.**[7] **G06F 17/30**

(52) **U.S. Cl.** **707/3**; 707/2; 707/4; 707/5; 707/6; 707/10; 705/5; 705/14; 705/26

(58) **Field of Search** 707/2, 4, 5, 100, 707/523, 3, 6, 10; 705/14, 26, 27, 5; 709/223, 228, 203, 219; 704/10; 713/171

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,640,553 A		6/1997	Schultz	707/5
5,675,819 A		10/1997	Schuetze	704/10
5,848,396 A		12/1998	Gerace	705/10
5,895,454 A		4/1999	Harrington	705/26
5,913,210 A	*	6/1999	Call	707/4
5,918,214 A		6/1999	Perkowski	705/27
5,960,429 A	*	9/1999	Peercy et al.	707/5
6,006,218 A	*	12/1999	Breese et al.	707/3
6,009,459 A		12/1999	Belfiore et al.	709/203
6,014,664 A		1/2000	Fagin et al.	707/5
6,064,979 A	*	5/2000	Perkowski	705/26
6,064,980 A	*	5/2000	Jacobi et al.	705/26
6,073,135 A		6/2000	Broder et al.	707/100
6,125,395 A		9/2000	Rosenberg et al.	709/228
6,144,958 A	*	11/2000	Ortega et al.	707/5
6,182,050 B1		1/2001	Ballard	705/14
6,230,153 B1	*	5/2001	Howard et al.	707/2
6,247,130 B1	*	6/2001	Fritsch	713/171
6,266,649 B1	*	7/2001	Linden et al.	705/26
6,401,118 B1	*	6/2002	Thomas	709/224
6,405,175 B1	*	6/2002	Ng	705/14
6,415,320 B1	*	7/2002	Hess et al.	709/219
6,418,434 B1	*	7/2002	Johnson et al.	707/5
6,421,675 B1	*	7/2002	Ryan et al.	707/100
6,460,072 B1	*	10/2002	Arnold et al.	709/203
6,484,149 B1	*	11/2002	Jammes et al.	705/26
6,535,896 B2	*	3/2003	Britton et al.	707/523

FOREIGN PATENT DOCUMENTS

GB	2 331 166 A	12/1999	

OTHER PUBLICATIONS

Kim et al., "Intelligent Information Recommend system on the Internet", IEEE, 1999, pp. 1–5.*
Seki et al., "User's behavior and URL analysis at EC sites", IEEE, Oct. 1999, pp. 87–92.*

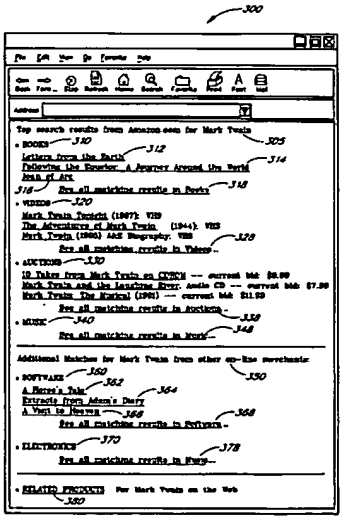
(List continued on next page.)

Primary Examiner—Thuy N. Pardo
(74) *Attorney, Agent, or Firm*—Knobbe, Martens, Olson & Bear LLP

(57) **ABSTRACT**

A search engine system assists users in locating web pages from which user-specified products can be purchased. Web pages located by a crawler program are scored, based on a set of rules, according to likelihood of including an online product offering. A query server accesses an index of the scored web pages to locate pages that are both responsive to a user's search query and likely to include a product offering. In one embodiment, the responsive web pages are listed on a composite search results page together with products that satisfy the query.

28 Claims, 9 Drawing Sheets



2. Second Analysis Stage

In practice, the vast majority of the web pages on the World Wide Web are not associated with product offerings, and as such their corresponding product scores are low. As shown in FIG. **5**, these web pages are excluded from the Product Spider database **147** by a filtering step **530**. The filter is simply a threshold number, preferably thirty, that the web page product score must equal or exceed to satisfy the filter. Web pages having a product score below thirty are discarded **532** as inappropriate for the Product Spider database **147**. Typically about 99% of all web pages in the World Wide Web are discarded in this manner. Those pages having product scores satisfying the filter criteria are retained. The corresponding URLs are submitted back **540** to the web crawler **160** for a second crawling stage **560**.

In other embodiments, such as those in which the index is also used to provide a general purpose web search engine, pages may be indexed without regard to their respective product scores. In still other embodiments, the filter comprises multiple ranges of product score values with predetermined minimum and maximum values. For example, four separate databases may be created for web pages having product score values of 20–40, 40–60, 60–80, and 80–100, respectively. In these latter embodiments the product scores may optionally be omitted from the respective databases.

If the Product Spider database is not being constructed for the first time, but rather is being updated, then the URLs from the existing database **147** are submitted **530** to the second crawling stage **560** as well. Duplication between the previous database submissions **550** and the latest web crawl submissions **540** are detected and removed (not shown).

The second crawling stage **560** shown in FIG. **5** typically requires substantially less time than the first crawling stage **510**, as the number of web pages involved is considerably smaller. The results of the second web crawling stage are passed through a second page analyzing stage **570**, wherein product scores are generated anew. In a second filtering stage **580**, pages failing to satisfy the filter are once again discarded **582**. Those pages satisfying the second filtering stage **580** are passed in step **590** to the index tool **164** for further processing.

The second filtering stage **580** preferably uses the same criteria as the first filtering stage **530**. In an alternative embodiment, the second filtering stage **580** may have either more or less discriminating criteria than the first filtering stage **530**.

3. Construction of the Product Spider Database

The pages retained after the second filtering stage **580** shown in FIG. **5** are passed to an indexing stage **590** wherein the index tool **164** creates the Product Spider database **147**, fully text indexed by keyword **166**. A given web page will contain multiple index keywords distributed throughout its text. The index tool **164** converts the information from a form organized by URL into a form organized by keyword. Schematically, the index tool **164** reorganizes the set of multiple pages ($Page_m$, where m=1 to M) containing multiple Keywords ($Word_n$, where n=1 to N) such that $Page_1$ ($\Sigma_n Word_n$), $Page_2$ ($\Sigma_n Word_n$), . . . , $Page_M(\Sigma_n Word_n)$ is converted into $Word_1(\Sigma_m Page_m)$, $Word_2$ $(\Sigma_m Page_m)$, . . . , $Word_N(\Sigma_m Page_m)$.

As shown in FIG. **1**, the database **147** includes, for each keyword **166**, one or more web page addresses **167** with corresponding titles **168**, squibs **169**, and product scores **170**. All of the product scores will necessarily equal or exceed thirty in the preferred embodiment due to the second filtering stage **580**.

The web page addresses **167** stored in the Product Spider database **147** are preferably "canonicalized" URLs. URLs often include one or more strings of characters appended to the addressing information that specify, for example, a particular user ID, session ID, or transaction ID. These characters are not needed for accessing the web page, and are thus preferably discarded, resulting in a "canonical" URL for inclusion in the Product Spider database **147**. Techniques for canonicalizing URLs are well known in the art.

The title **168** entry of the database **147** is preferably duplicated directly from the title used for the web page, as identified by the appropriate HTML tags. If a web page has an inappropriate title, or is missing a title, a new title is inserted into the database **147** as needed on a case by case basis.

The squib **169** entry of the database **148** is generated automatically by the index tool **164**. The squib corresponds to the initial series of words on a web page, up to a preset number of characters set at about two-hundred. In another embodiment, the squib displays relevant text extracted from the web page corresponding to the products offered for sale on the web page.

The process illustrated in FIG. **5** may be used to update the Product Spider database **147** as often as desired. In a preferred embodiment, the Product Spider database **147** is updated every week, more preferably the database is updated every three or four days, and even more preferably it is updated every day.

As indicated above, the Product Spider database **147** may alternatively be constructed without storing the product scores for each page. In one embodiment, for example, the database comprises only pages having a product score satisfying predetermined criteria, for example, requiring the product score to equal or exceed thirty (as in the filtering steps **530**, **580** of FIG. **5**). In another alternative embodiment, the database comprises multiple indexed tables created without storing the product scores, wherein each table is constructed from web pages having a product score satisfying unique criteria, for example, four separate indexed tables containing pages having product scores from 20–40, 40 –60, 60–80 and 80–100, respectively.

In another embodiment, the Product Spider database **147** consists of multiple indexed tables, wherein each table is constructed from web pages that are distinguishable on the basis of some aspect of product offerings (ascertained from parsing the web pages) unrelated to product scores. In one embodiment, for example, the database **147** consists of separate tables for different categories of goods (e.g., books, music, videos, electronics, software, and toys). In another embodiment, a separate table is used for products unsuitable for children. In still another embodiment, different tables are constructed for web sites written in different languages (English, Japanese, German, etc.). In yet another embodiment, different tables are constructed for on-line and off-line product offerings. Under these embodiments, the page analyzer steps **520**, **570** include searching for character strings judged to be associated with the various predefined categories.

By constructing the Product Spider database **147** out of different tables having distinguishing characteristics, or retaining the equivalent information within one big table, the user is capable of conducting a more refined search within the Product Spider database **147**. In one embodiment, for example, the Related Products hypertext link **380** is replaced by a pulldown menu comprising different categories corresponding to the distinctions retained within the Product Spider database **147** (e.g., books, music, video, and toys categories, on-line versus off-line offerings, goods versus services, etc.).